

LLM-based Fusion of Multi-modal Features for Commercial Memorability Prediction

Aleksandar Pramov

Georgia Institute of Technology, USA

Abstract

This paper addresses the prediction of commercial (brand) memorability as part of “Subtask 2: Commercial/Ad Memorability” within the “Memorability: Predicting movie and commercial memorability” task at the MediaEval 2025 workshop competition. We propose a multimodal fusion system with a Gemma-3 LLM backbone that integrates pre-computed visual (ViT) and textual (E5) features by multi-modal projections. The model is adapted using Low-Rank Adaptation (LoRA). A heavily-tuned ensemble of gradient boosted trees serves as a baseline. A key contribution is the use of LLM-generated rationale prompts, grounded in expert-derived aspects of memorability, to guide the fusion model. The results demonstrate that the LLM-based system exhibits greater robustness and generalization performance on the final test set, compared to the baseline.

The paper’s codebase can be found at <https://github.com/dsgt-arc/mediaeval-2025-memorability>

1. Introduction

Video memorability plays a central role in various applications such as marketing and advertisement, film-making, and higher education. Modeling video memorability is particularly challenging because it requires integrating multi-channel data (visual, audio, and text) to predict a latent, unobservable, subjective characteristic of the data.

The “Memorability: Predicting movie and commercial memorability” task at MediaEval 2025 workshop competition undertakes to study methods that can both predict memorability, as well as shed light on the drivers behind both movie and commercial memorability [1]. This paper deals in particular with the latter question (“Subtask 2: Commercial/Ad Memorability”) by modeling the memorability score of commercial videos (from the financial industry) based on video and text features, as well as their brand memorability (i.e., how well a brand is remembered from a video). To that end, this year’s competition dataset includes text and pre-computed video features of 424 curated YouTube commercials from financial institutions. The provided data dictates the need to integrate multi-modal features for predictive modeling: numeric (e.g., video engagement metrics), text (e.g., video titles and subtitles), and visual embeddings. One important modeling limitation was the lack of access to the raw videos, making the usage of off-the-shelf modern multimodal LLMs difficult [2]. The approach undertaken here thus focuses on feature integration by applying both a gradient boosting model as a baseline and leveraging an LLM for the multimodal fusion of video metadata, textual embeddings, and visual features [3].


MediaEval’25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

✉ apramov3@gatech.edu (A. Pramov)

🆔 <https://orcid.org/0009-0005-9049-1337> (A. Pramov)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Work

Previous works in video memorability prediction often relied on sophisticated feature engineering, integrating components such as deep visual features, textual semantics, and spatio-temporal information to enhance predictive capabilities [4, 5]. More recently, the field has embraced leveraging the vast world knowledge and complex reasoning abilities of Multimodal Large Language Models (MM-LLMs), such as Gemma and Qwen-VL, with fine-tuning techniques such as LoRA, for tasks like human perception analysis and long-term advertisement memorability [6, 3, 2, 7, 8].

3. Approach

3.1. Data manipulation & Evaluation setup

The training dataset for MediaEval’s 2025 Subtask 2 edition consists of a small sample of 339 videos for training and 85 videos for evaluation, sampled from the VIDEM Dataset [9] - this was the only dataset used in our work. Moreover, multiple videos are from the same channel on very similar topics, e.g., videos on company earnings calls from different quarters within the same year. To that end, we employ two data manipulation steps on our training sample of 339 videos: First, we break down the biggest channel (“Goldman Sachs”, with 23% of all the videos) into three “subchannels”, based on a k-medoids clustering algorithm on the embedded¹ video titles. This subgrouping creates a more balanced distribution of the channels (the largest one being 37). Second, we create a nested 5-fold, grouped (by channel name) & stratified (by 5 quantiles of the target variable) grouping of the data.

The aim of this nested, stratified and grouped evaluation setup is to estimate the generalization error in a more robust way, while guarding against information leakage. We end up having 5 folds in the outer loop, each with 5 inner train/validation splits that will be used to elicit hyperparameters, early stopping and/or estimate the validation error.²

3.2. Model features

For the modeling, we investigated the following provided (or derived) features as input to the models, the detailed integration of which is discussed in the next section:

- *Numerical metadata*, provided by the organizers.
- *E5-base-v2 Embeddings* of the *subtitles* (chunked and pooled), *titles*, and *descriptions*.
- Pre-computed *video embeddings*, provided by the organizers.
- *Subtitle summaries* of up to 1024 tokens, generated by `gemma3-4b-it-qat`, used both as *E5-base-v2* embeddings and as text for model prompts.
- Fold-aware few-shot `gemma3-4b-it-qat` generated (*brand*) *memorability text rationales*, based on the subtitles of the videos. The aim of the prompt was to mimic an expert system to evaluate (qualitatively) the (brand) memorability of the video, based on the subtitles only, along key characteristics such as brand integration, clarity of brand messaging, semantic richness, novelty and others.³ The generated rationales were then either used

¹<https://huggingface.co/intfloat/e5-base-v2>

²Throughout the rest of this manuscript, “CV SRCC” refers to the averaged test error across the 5 outer folds. The performance on the 85 held-out videos from the competition is referred to as “Test SRCC.”

³The few-shot prompting here did not involve providing examples of full text of rationales, but there was fold-adaptive guidance with a few-shot examples of 2-3 (brand) memorability scores of the respective inner fold. These

as E5-base-v2 embeddings, or in text form, to be a part of a prompt in one of the memorability predictive systems.

3.3. Modeling

We investigated two modeling architectures with (brand) memorability as a univariate target variable. The first, a baseline histogram gradient boosted tree model (**HGBT**), took as input the various text and visual embeddings, as well as the numerical metadata. This approach naturally has an overwhelming number of features for a very limited sample; thus, strong PCA reduction for each of the individual text and image feature streams was employed as a pre-processing step. The hyperparameters of the HGBT models were tuned on the inner folds using Optuna to optimize for Spearman correlation, and the final model for each outer fold was evaluated on the outer test split.

The second approach delves into the complexity of joint textual and visual modeling and closely follows a previously outlined approach by Esteban et al., of using instruction-tuned LLMs as multi-modal feature integrators [3, 10, 11]. Our implementation, which we call **Gemma Fusion**, uses a gemma-3-4b-it⁴ model as its backbone. We augment the model’s standard text-based input with external feature streams that are projected into the LLM’s embedding space as via separate trainable linear projectors, with an early fusion step at the embedding level. The textual prompt is constructed from the video’s title and the aforementioned LLM-generated text, which is either a summary of the subtitles or a qualitative rationale on the video’s memorability.

In parallel, external features—including E5 embeddings for subtitles, title, and description, as well as pre-computed visual blocks through the linear projector, effectively adapting the (visual) input features to be considered as input tokens for the LLM. The fused unified sequence is then fed to the Gemma backbone. The last hidden state from the transformer blocks is (mean or attention pooled) and passed through an MLP head producing a (brand) memorability prediction. The MLP and the projectors are trained, optimizing using an equally weighted composite loss function that is a weighted average of Mean Absolute Error (MAE) and a correlation coefficient [3]. The gemma backbone is either frozen or Low-Rank Adaptation (LoRA) is performed on the query, key, value, and output projections in the attention layers, as well as the gate, up, and down projections of the feed-forward layers.

4. Results and Analysis

The final results of our experiments, summarizing both the 5-fold cross-validation (CV) and the official competition test set performance, are presented in Table 1. Only the ViT embeddings were used, as analyses showed that including the other provided embeddings did not yield an added value to our model. Given the very small size of the sample, we decided to pick a visual embedding model, the performance of which on such task is grounded in previous literature [3].

One key highlight is that, despite the extensive efforts, overfitting did occur quite substantially for the HGBT models. The performance on the competition dataset collapsed, compared to the CV procedure. This suggests that while the feature engineering was effective on the development data, the model ultimately overfit to its specific characteristics.

examples were chosen to be of those videos, that were the nearest neighbors by cosine similarity in the (pooled and averaged) embedding space of subtitles, titles and description.

⁴<https://huggingface.co/google/gemma-3-4b-it>

Table 1**Cascaded ablation and final submission results**

Comparing the mean of *outer fold* 5-fold cross-validation (CV) scores with official competition test set scores. Final submitted models are indicated by a result in the Test SRCC column. **SumEmb**=E5(Summaries), **RatEmb**=E5(Rationales) embeddings. **Attn./Mean**=Pooling. **P**=Prompt content (Rat.=Rationales, Sum.=Summaries). See main text for detailed information the models and the nested CV construction.

Target	Model	Config.	CV SRCC	CV RMSE	Test SRCC
Brand Memorability	HGBT	Base (E5(Text)+Numeric)	0.0021	0.1662	-
		Ablation (Base + ViT)	0.0304	0.1569	-
		+ SumEmb	0.1181	0.1547	0.019
		+ RatEmb	0.0504	0.1574	-
	Gemma Fusion	Base (P: Rat)	0.0588	0.1701	-
		Ablation (+E5(Text)+ViT, Attn.)	0.0223	0.1534	-
		Final (E5(Text)+ViT, Mean, P: Rat.)	0.1569	0.1506	0.122
		Final (E5(Text)+ViT, Mean, P: Sum.)	0.1627	0.1515	0.112
Memorability Score	HGBT	Base (E5(Text)+Numeric)	0.1190	0.1490	-
		Ablation (Base + ViT)	0.1577	0.1493	-
		+ SumEmb	0.1955	0.1436	-
		+ RatEmb	0.2402	0.1425	0.089
	Gemma Fusion	Final (E5(Text)+ViT, Mean, P: Rat.)	0.1165	0.1580	0.018
		Final (E5(Text)+ViT, Mean, P: Sum.)	0.0849	0.2158	0.131

In contrast, the Gemma Fusion model exhibited greater robustness and superior generalization - the LLM-based feature integration approach exhibited much better performance in three of the four submissions. Using LoRA proved beneficial in multiple ablations (not all shown here) and thus all final Gemma employed LoRA with a rank of 32, an alpha of 32, and a dropout of 0.15, applied to both attention projections and feed-forward layers. The best-performing setups consistently used mean pooling and included the ViT visual block and all three E5 text streams (subtitles, title, and description). Notably, the performance of the prompt content was target-dependent. On the final test set, Brand Memorability benefited more from rationales as part of the prompt (0.122 vs. 0.112 SRCC), while Memorability Score performed substantially better using the subtitle summary (0.131 vs. 0.018 SRCC).

5. Discussion and Outlook

The small sample size proved to be a significant challenge for the task. On the one hand, a simpler model like HGBT still overfit and did not prove to be a good baseline on the competition dataset. On the other hand, our LLM-multimodal fusion approach exhibited much more robust performance but still indicates issues with the training stability, as evidenced by the collapse of the rationales run for the memorability score (from 0.1165 on the CV SRCC to 0.018 on the test). Nonetheless, the LLM-multimodal fusion approach managed to lift the the baseline performance both in absolute terms and in terms of stability of the system. In addition, the generated LLM rationales introduced a novel and alternative way to prompt such a fused model, which did lift the performance for the brand memorability.

Future work will involve the addition of additional datasets for training, e.g. *memento10k* which should improve the overall stability, and possibly performance of the model. In addition, crafting better, (brand) memorability-expert oriented prompts for the model could be another promising avenue for further research. Lastly, throughout the whole analysis we did not leverage specific domain information - models which are fine-tuned on textual data from financial domain could be a better choice and have the potential to bring added value.

Acknowledgments

We thank the DS@GT team for providing valuable comments and suggestions. This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA.

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 2.5 in order to perform grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

- [1] I. Martín-Fernández, M. G. Constantin, C.-H. Demarty, M. Gil-Martín, S. Halder, B. Ionescu, A. Matran-Fernandez, R. Savran Kiziltepe, A. García Seco de Herrera, Overview of the mediaeval 2025 predicting movie and commercial memorability task, in: Proc. of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.
- [2] I. Martín-Fernández, S. Esteban-Romero, F. Fernández-Martínez, M. Gil-Martín, Parameter-efficient adaptation of large vision–language models for video memorability prediction, *Sensors (Basel, Switzerland)* 25 (2025) 1661.
- [3] S. Esteban-Romero, I. Martín-Fernández, M. Gil-Martín, D. Griol-Barres, Z. Callejas-Carrión, F. Fernández-Martínez, Llm-driven multimodal fusion for human perception analysis, in: Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor, 2024, pp. 45–51.
- [4] S. Shekhar, D. Singal, H. Singh, M. Kedia, A. Shetty, Show and recall: Learning what makes videos memorable, in: Proceedings of the IEEE international conference on computer vision workshops, 2017, pp. 2730–2739.
- [5] I. Martín-Fernández, S. Esteban-Romero, J. Bellver-Soler, F. Fernández-Martínez, M. Gil-Martín, Larger encoders, smaller regressors: Exploring label dimensionality reduction and multimodal large language models as feature extractors for predicting social perception, in: Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor, 2024, pp. 20–27.
- [6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., *ICLR* 1 (2022) 3.
- [7] S. Harini, S. Singh, Y. K. Singla, A. Bhattacharyya, V. Baths, C. Chen, R. R. Shah, B. Krishnamurthy, Long-term ad memorability: Understanding & generating memorable ads, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 5707–5718.
- [8] I. Martín-Fernández, R. Kleinlein, C. Luna-Jiménez, M. Gil-Martín, F. Fernández-Martínez, Video memorability prediction from jointly-learned semantic and visual features, in: Proceedings of the 20th international conference on content-based multimedia indexing, 2023, pp. 178–182.
- [9] R. S. Kiziltepe, S. Sahab, R. V. Santana, F. Doctor, K. Paterson, D. Hunstone, A. G. S. de Herrera, VIDEM: VIDEO effectiveness and memorability dataset, in: I. Rojas, G. Joya, A. Catala (Eds.), *Advances in Computational Intelligence*, Springer Nature Switzerland, Cham, 2025, pp. 41–54.
- [10] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, *Advances in neural information processing systems* 36 (2023) 34892–34916.
- [11] G. Verma, M. Choi, K. Sharma, J. Watson-Daniels, S. Oh, S. Kumar, Cross-modal projection in multimodal llms doesn't really project visual attributes to textual space, *arXiv preprint arXiv:2402.16832* (2024).